DOCUMENT RESUME

ED 467 822 TM 034 366

AUTHOR	Glas, C. A. W.
TITLE	CUSUM Statistics for Large Item Banks: Computation of Standard Errors. Law School Admission Council Computerized Testing Report. LSAC Research Report Series.
INSTITUTION	Law School Admission Council, Princeton, NJ.
REPORT NO	LSAC-R-98-11
PUB DATE	2001-09-00
NOTE	16p.; For the previous study of online calibration in computerized assessment, see ED 421 549.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	*Adaptive Testing; *Computer Assisted Testing; *Error of Measurement; Estimation (Mathematics); Goodness of Fit; Item Banks; Item Response Theory; Online Systems; Simulation; *Test Items
IDENTIFIERS	*Calibration

ABSTRACT

In a previous study (1998), how to evaluate whether adaptive testing data used for online calibration sufficiently fit the item response model used by C. Glas was studied. Three approaches were suggested, based on a Lagrange multiplier (LM) statistic, a Wald statistic, and a cumulative sum (CUMSUM) statistic respectively. For all these methods, the asymptotic variance of the parameter estimates has to be approximated. In the previous study, standard errors were computed using an observed Fisher information matrix. However, when the number of items in each bank becomes very large, manipulating complete information matrices becomes quite difficult. This study investigates the extent to which standard errors can be computed using the diagonal of information matrices only, and how the CUMSUM procedure must be tuned to this alternative approach. Simulation studies showed that the asymptotic standard errors are underestimated by the block-diagonal approach but that the magnitude of the bias in the standard errors was relatively small. It was also shown that the power of the statistical test based on a CUMSUM statistic using these approximated standard errors is well under control. (Author/SLD)

Reproductions supplied by EDRS are the best that can be made

ED 467 822

LSAC RESEARCH REPORT SERIES

TM

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

J. VASELECK

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

■ CUSUM Statistics for Large Item Banks: Computation of Standard Errors

C. A. W. Glas
University of Twente

■ Law School Admission Council
Computerized Testing Report 98-11
September 2001

TM034366

BEST COPY AVAILABLE

The Law School Admission Council is a nonprofit corporation that provides services to the legal education community. Its members are 199 law schools in the United States and Canada.

Copyright © 2001 by Law School Admission Council, Inc.

All rights reserved. No part of this report may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without permission of the publisher. For information, write: Communications, Law School Admission Council, Box 40, 661 Penn Street, Newtown, PA 18940-0040.

LSAT® and the Law Services logo are registered marks of the Law School Admission Council, Inc.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in these reports are those of the authors and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary 1

Abstract 1

Introduction 1

The CUSUM Statistic 2

Computation of Standard Errors 3

A Simulated Comparison of Approximations of Standard Errors 4

The Power of the CUSUM Statistic 9

Conclusion 11

References 11

Executive Summary

In a computer-based testing (CBT) or computerized adaptive testing (CAT) environment, the process of statistically calibrating the pool of test items may consist of the following two stages:

1. *Pretesting stage.* In this stage, subsets of items are administered to subsets of test takers in a series of pretest sessions, and an item response theory (IRT) model is fitted to the response data to obtain empirical estimates of such item properties as their difficulty, discriminating power, and liability to guessing.

2. *Online stage.* In this stage, response data are gathered in an operational computerized testing environment and used to estimate the values of the item parameters. An essential feature of online calibration is that the data are gathered sequentially and the item parameter estimates are improved in a gradual manner.

In a previous report by Glas, statistical tests for detecting unwanted differences between results from pretesting and online calibration were studied. These tests were based on three classes of statistics: Lagrange multiplier, Wald, and CUSUM statistics. For each statistic the standard errors of the parameter estimates have to be approximated. In the previous report, these standard errors were computed for the 2-parameter logistic (2-PL) and 3-parameter logistic (3-PL) models with fixed values for the guessing parameter and an approximate Fisher information matrix. However, when the number of items in the item bank becomes large, inversion of information matrices becomes a very time-demanding operation. Therefore, in the present report, the use of a block-diagonal approximation to the Fisher information matrix is investigated.

In the 2-PL model and the 3-PL model with fixed guessing parameter, every item is characterized by two unknown parameters: the discrimination and the difficulty parameter. In the diagonal approach, item information is approximated by a two-by-two matrix with diagonal entries for information on both item parameters and off-diagonal entries to account for the common part in their information. Information across items is not taken into account. Using simulation studies, it was shown that the asymptotic standard errors are underestimated by the block-diagonal approach but that the magnitude of the bias in the standard errors was relatively small. Further, it was shown that the power of the statistical test based on a CUSUM statistic using these approximated standard errors is well under control.

When employing the CUSUM test in practice, it is suggested that the CUSUM statistic be tuned to the application by running a simulation study with a test administration design and item parameter values similar to those in the real application. The results of the simulation study can then be used to determine how underestimated standard errors can be translated into threshold values for the CUSUM statistic that guarantee a test with acceptable power.

Abstract

In a previous report it was investigated how to evaluate whether adaptive testing data used for online calibration sufficiently fit the item response model used by Glas. Three approaches were suggested, based on a Lagrange multiplier (LM) statistic, a Wald statistic, and a cumulative sum (CUSUM) statistic, respectively. For all these methods, the asymptotic variance of the parameter estimates has to be approximated. In the previous report, standard errors were computed using an approximate observed Fisher information matrix. However, when the number of items in the item bank becomes very large, manipulating complete information matrices becomes quite difficult. Therefore, in the present report it is investigated to what extent standard errors can be computed using the diagonal of information matrices only, and how the CUSUM procedure must be tuned to this alternative approach.

Introduction

In a computer-based testing (CBT) or computerized adaptive testing (CAT) environment, the calibration process may consist of two stages:

(1) *The Pretesting stage.* In this stage, subsets of items are administered to subsets of respondents in a series of pretest sessions, and an item response model is fitted to the data to obtain item parameter estimates to support computerized test administration.

(2) *The online stage.* In this stage, data are gathered in a computerized assessment environment and incoming data are used for further parameter estimation.

BEST COPY AVAILABLE

Evaluating differences in model fit of data emanating from the pretesting and online stage can be done using Lagrange multiplier (LM) statistics, Wald statistics, and CUSUM statistics (Glas, 1998c). For all these approaches, the asymptotic variance of the parameter estimates has to be approximated. In a previous report, standard errors were computed using an approximate observed Fisher information matrix. However, when the number of items in the item bank becomes very large, inversion of information matrices becomes quite difficult. Therefore, in the present report it is investigated to what extent standard errors can be computed using the diagonal of information matrices only.

This paper is organized as follows: In the section that follows, the principle of the CUSUM statistic will be sketched. Next, methods for computation of standard errors will be discussed and these methods will be evaluated using a number of simulation studies. The power of the CUSUM statistic using a diagonal approximation to the observed information matrix will then be studied using artificial data. Finally, some conclusions and suggestions for further research will be formulated.

The CUSUM Statistic

In this report, the 3-parameter logistic (3-PL) model will be used for describing response behavior. The motivation is that CAT usually employs dichotomously scored selected response items. A problem is that guessing may be prominent in the calibration phase, while it may occur less frequently in the online phase, because here the items are tailored to the ability level of the respondents. Consequently, estimation of the guessing parameter in the online phase may be very difficult. Therefore, in the present paper it will be assumed that the guessing parameter is fixed to some plausible constant, say, to the opposite of the number of response alternatives available.

So, consider dichotomous items where responses of persons labeled n to items labeled i are coded $x_{ni} = 0$, and $x_{ni} = 1$. The probability of a correct response is given by

$$\begin{aligned}\phi_i(\theta_n) &= Pr(X_{ni} = 1 | \theta_n, \alpha_i, \beta_i, \gamma_i) \\ &= \gamma_i + (1-\gamma_i)\psi_i(\theta_n) \\ &= \gamma_i + (1-\gamma_i)\frac{\exp(\alpha_i\theta_n - \beta_i)}{1+\exp(\alpha_i\theta_n - \beta_i)},\end{aligned} \tag{1}$$

where θ_n is the ability parameter of person n and α_i, β_i and γ_i are the discrimination, difficulty, and guessing parameter of item i, respectively. The CUSUM chart is an instrument of statistical quality control used for detecting small changes in product features during the production process.

The CUSUM chart is used in a sequential statistical test, where the null-hypothesis of no change is never accepted (Veerkamp, 1996). In the present case, the procedure starts with parameter estimates α_{i0} and β_{i0} obtained in the pretest phase. Then, for new batches of respondents $g = 1, ..., G$ partaking in the tailored test phase, the alternative hypothesis entails that the item is becoming more easy and is losing its discriminating power. So the null-hypothesis is $\alpha_{ig} - \alpha_{i0} \geq 0$ and $\beta_{ig} - \beta_{i0} \geq 0$, for $g = 1, ..., G$. A cumulative sum chart will be based on the quantity

$$S_i(g) = \max\left\{S_i(g-1) + \frac{\alpha_{i0} - \alpha_{ig}}{Se(\alpha_{ig} - \alpha_{i0})} + \frac{\beta_{i0} - \beta_{ig}}{Se(\beta_{i0} - \beta_{ig} | \alpha_{i0} - \alpha_{ig})} - k, 0\right\} \tag{2}$$

where

$$Se(\alpha_{ig} - \alpha_{i0}) = \sigma_\alpha$$

and

$$Se(\beta_{i0} - \beta_{ig} | \alpha_{i0} - \alpha_{ig}) = \sqrt{\sigma_\beta^2 - \sigma_{\alpha,\beta}^2 / \sigma_\alpha^2},$$

with

$$\sigma_{\alpha}^{2}, \sigma_{\beta}^{2} \text{and} \sigma_{\alpha,\beta}$$

the appropriate elements of the covariance matrix of the parameter estimates. Further, k_i is a reference value. The CUSUM chart starts with $S_i(0) = 0$ and the null-hypothesis is rejected as soon as $S_i(j) > h$, where h is some constant threshold value. The choice of the constants k and h determines the power of the procedure. In the case of the Rasch model, where the null-hypothesis is $\beta_{ix} - \beta_{i0} \geq 0$, and the term involving the discrimination indices is lacking from Equation 2, Veerkamp (1996) successfully uses $k = 1/2$ and $h = 5$. This choice was motivated by the consideration that this setup has good power against the alternative hypothesis of a normalized shift in item difficulty of approximately one standard deviation. In the present case one extra normalized decision variable is employed, that is, the variable involving the discrimination indices. To have power against a shift of one standard deviation of both normalized decision variables in the direction of the alternative hypothesis, a value $k = 1$ will be used below. The value $h = 5$ will not be changed.

Computation of Standard Errors

Let the item administration variable d_{ni} take the value one if the item was administered to n and zero if this was not the case. If $d_{ni} = 0$ it will be assumed that $x_{ni} = c$, where c is some arbitrary constant. Let x_n and d_n be the response pattern and the item administration vector of respondent, respectively. In this report, the framework of marginal maximum likelihood (MML) will be adopted. So, it will be assumed that ability is normally distributed. Let $g(.;\mu,\sigma)$ be the density of θ_n. Further, ξ is a vector of all parameters in the model, that is, $\xi' = (\alpha', \beta', \gamma', \mu, \sigma)$. The log-likelihood to be maximized can be written as

$$\begin{aligned} \ln L(\xi; X, D) &= \sum_n \ln p(x_n \mid d_n; \xi) \\ &= \sum_n \ln \int p(x_n \mid d_n, \theta_n, \alpha, \beta, \gamma) g(\theta_n \mid \mu, \sigma) \partial\theta_n , \end{aligned} \tag{3}$$

where X stands for the data matrix and D stands for the design matrix.

The asymptotic covariance matrix of the parameter estimates can be computed by inverting the so-called observed Fisher information matrix (see Efron & Hinkley, 1978):

$$H(\xi, \xi) = -\frac{\partial^2 \ln L(\xi; X, D)}{\partial\xi \partial\xi'} . \tag{4}$$

Using an identity by Louis (1982, also see, Glas, 1992, 1998a, 1998b, 1998c), the observed Fisher information matrix can be written as a sum over respondents of terms

$$H_n(\xi, \xi) = -E(B_n(\xi, \xi) \mid x_n, d_n, \xi) - Cov\left(b_n(\xi) b_n(\xi)' \mid x_n, d, \xi\right), \tag{5}$$

with

$$B_n(\xi, \xi) = \frac{\partial^2 \ln p(x_n, \theta_n \mid d_n; \xi)}{\partial\xi \partial\xi'} ,$$

$$Cov\left(b_n(\xi) b_n(\xi)'\right) = E\left(b_n(\xi) b_n(\xi)' \mid x_n, d, \xi\right) + h_n(\xi) h_n(\xi)' ,$$

$$h_n(\xi) = \frac{\partial}{\partial\xi} \ln p(x_n \mid d_n; \xi) = E(b_n(\xi) \mid x_n, d_n, \xi) ,$$

BEST COPY AVAILABLE

and

$$b_n(\xi) = \frac{\partial}{\partial \xi} \ln p(x_n, \theta_n \mid d_n; \xi).$$

For the 3-PL the exact expressions for the second order derivatives may become rather complicated. Therefore, one may use the approximation

$$H(\xi, \xi) = \sum_n h_n(\xi) h_n(\xi)' \tag{6}$$

(also, see Mislevy, 1986). This approximation is based on the following argument. To keep the argument simple, the design vector is temporarily discarded. Let $\{x\}$ be the set of all possible response patterns. Differentiating both sides of the equation

$$\sum_{\{x\}} \int p(x, \theta; \xi) \partial\theta = 1$$

twice with respect to ξ results in the well known identity

$$\sum_{\{x\}} \int \frac{\partial^2 \ln p(x, \theta; \xi)}{\partial \xi \partial \xi'} p(x, \theta, \xi) \partial\theta = -\sum_{\{x\}} \int \left\{ \frac{\partial \ln p(x, \theta; \xi)}{\partial \xi} \right\} \left\{ \frac{\partial \ln p(x, \theta; \xi)}{\partial \xi'} \right\} p(x, \theta, \xi) \partial\theta$$

which can also be written as

$$\sum_{\{x\}} \int \frac{\partial^2 \ln p(x, \theta; \xi)}{\partial \xi \partial \xi'} p(\theta \mid x; \xi) \partial\theta \, p(x; \xi) = -\sum_{\{x\}} \int \left\{ \frac{\partial \ln p(x, \theta; \xi)}{\partial \xi} \right\} \left\{ \frac{\partial \ln p(x, \theta; \xi)}{\partial \xi'} \right\} p(\theta \mid x; \xi) \partial\theta \, p(x; \xi).$$

Replacing the sum over the distribution of $\{x\}$ by a sum over observed response patterns results

$$\sum_n E(B_n(\xi, \xi) \mid x_n, d_n, \xi) = \sum_n E\left(b_n(\xi) b_n(\xi)' \mid x_n, d, \xi \right),$$

and Equation 4 can be approximated by Equation 6.

A Simulated Comparison of Approximations of Standard Errors

Various approximations to the standard errors were completed in three simulation studies, concerning the 1-, 2- and 3-PL model, respectively. For all studies, data were simulated using a complete design where all simulated test takers responded to all items. Since this section does not represent an adaptive testing situation, adaptive testing will be invoked in the next section. The first simulation study concerned approximations to standard errors for the 1-PL model. For sample sizes of 250, 1,000, and 4,000 simulated test takers, data were generated for test models consisting of 5, 10, 20, and 40 items. Data were generated using the 1-, 2- and 3-PL models. The item parameters are given in Table 1. The second and third column pertain to studies with 5 items, the fourth and fifth column pertain to studies with 10 items, and the last two columns pertain to studies with 20 items. In 40-item-studies, the parameter vector of the 20-item-studies was duplicated. Table 1 gives the item parameters for the 2-PL model; for the l-PL model, all discrimination indices α were set equal to one, while the item difficulties β were left unchanged; for the 3-PL model α and β were equal to the values of the 1 -PL model and the guessing parameter γ was equal to 0.20 for all items. Notice that for the 2-PL model, the item parameters are chosen in such a way that items with relative extreme absolute difficulty have lower discrimination indices. This was done to ensure the stability of the parameter estimates.

TABLE 1
Item parameters for simulation studies

i	α	β	α	β	α	β
1	0.8	-0.8	0.7	-0.8	0.7	-0.9
2	0.9	-0.4	0.8	-0.6	0.8	-0.8
3	1.0	0.0	0.9	-0.4	0.9	-0.7
4	0.9	0.4	1.0	0.0	1.0	-0.6
5	0.8	0.8	1.2	0.0	1.1	-0.5
6			0.8	0.0	1.2	-0.4
7			1.0	0.4	1.3	-0.3
8			0.9	0.6	0.8	0.0
9			0.8	0.8	1.0	-0.1
10			0.7	1.0	1.0	0.0
11					1.0	0.1
12					1.0	0.0
13					0.8	0.3
14					1.3	0.4
15					1.2	0.5
16					1.1	0.6
17					1.0	0.7
18					0.9	0.8
19					0.8	0.9
20					0.7	1.0

Consider the overview of Table 2. In the first two columns of this table, labeled N and K, the sample sizes and test lengths are given. For the time being, ignore the seven bottom lines of the table, they contain an overview that will be returned to later. Since data were generated using the 1-, 2- and 3-PL, the third column, labeled PL is related to the model used for data generation. Every line in the table contains the results of a specific combination of sample size, test length, and data generation model. For each line, 10 replications were made. Each replication consisted of generating a data set, estimating the parameters of the l-PL model, and computing various standard errors. These were computed as the inverse of the complete information matrix defined by Equation 5, its diagonal, the approximate matrix defined by Equation 6, its diagonal, and $\sum_n E\left(B_n\left(\xi,\xi\right)\mid x_n, d_n, \xi\right)$. This last approximation is motivated by the fact that these second-order derivatives are readily available since they are usually computed in the M-step of the EM-algorithm. It can easily be verified that the off-diagonal entries of this matrix are all equal to zero, so this approach also results in a diagonal approximation. Under the heading Diagonal Observed Info Matrix the mean over replications and parameters of the standard error obtained using the diagonal of the observed information matrix relative to the standard error obtained using the complete observed information matrix is displayed. The entries under Ccomplete Approximate Matrix, Diagonal Approximate Matrix, and M-step Approximate Matrix, are similar means obtained using the approximate information matrix defined by Equation 6, its diagonal, and the matrix of second order derivatives of the M-step, respectively.

TABLE 2

Approximation of standard errors for the 1-PL model magnitude relative to standard error, computed using complete observed information matrix

N	K	PL	Diagonal Observed Info Matrix	Complete Approximate Matrix	Diagonal Approximate Matrix	M-Step Approximate Matrix
250	5	1	0.953	0.996	0.953	0.902
		2	0.964	0.998	0.964	0.920
		3	0.963	0.991	0.963	0.924
	10	1	0.917	0.981	0.917	0.884
		2	0.939	0.979	0.939	0.909
		3	0.949	0.975	0.949	0.924
	20	1	0.904	0.968	0.904	0.885
		2	0.900	0.955	0.899	0.881
		3	0.917	0.953	0.916	0.900
	40	1	0.885	0.974	0.887	0.877
		2	0.878	0.972	0.880	0.869
		3	0.858	0.901	0.858	0.849
1,000	5	1	0.958	1.002	0.958	0.907
		2	0.966	1.002	0.966	0.920
		3	0.974	1.000	0.974	0.935
	10	1	0.942	0.997	0.941	0.909
		2	0.954	0.994	0.954	0.925
		3	0.962	0.994	0.962	0.935
	20	1	0.928	0.990	0.928	0.909
		2	0.924	0.988	0.924	0.905
		3	0.947	0.987	0.947	0.930
	40	1	0.947	1.021	0.949	0.938
		2	0.939	1.007	0.941	0.930
		3	0.929	0.978	0.929	0.919
4,000	5	1	0.963	1.003	0.963	0.913
		2	0.970	1.003	0.970	0.925
		3	0.977	1.001	0.977	0.939
	10	1	0.945	1.000	0.945	0.912
		2	0.954	0.998	0.954	0.924
		3	0.967	0.998	0.967	0.939
	20	1	0.932	1.001	0.933	0.913
		2	0.930	0.994	0.930	0.911
		3	0.956	0.995	0.956	0.938
	40	1	0.962	1.044	0.965	0.954
		2	0.954	1.031	0.956	0.945
		3	0.945	0.992	0.945	0.935
250	*	*	0.919	0.970	0.919	0.894
1,000	*	*	0.948	0.997	0.948	0.922
4,000	*	*	0.955	1.005	0.955	0.929
*	5	*	0.965	1.000	0.965	0.921
*	10	*	0.948	0.991	0.948	0.918
*	20	*	0.926	0.981	0.926	0.908
*	40	*	0.922	0.991	0.923	0.913

It can be seen that Equations 5 and 6 produce comparable results. This is in accordance with previous results (Glas, 1998b). Further, the diagonal approximations and the M-step approximation systematically underestimate standard errors, and the last approximation is least precise. However, overall the bias is relatively small. The seven last rows of the table give an indication of the main effect of sample size and test length. For instance, the row with entry 250 under N and two stars under K and PL gives the results for sample size 250 summed over all four test lengths and all three models. In the same manner, the last row contains the results for test length 40 summed over all three sample sizes and all three models. It can be seen that the precision of the diagonal approximations increases with sample size and decreases with test length.

In Table 3, the results for the standard errors of the 2-PL model are summarized. This table has a similar format to the previous table—the columns labeled α contain the results for the discrimination parameters, and the columns labeled β contain the results for the difficulty parameters. To account for the fact that in the 2-PL model every item is associated with two item parameters, the diagonal approaches are changed to block-diagonal approaches, in the sense that item information is now approximated by a two-by-two matrix with diagonal entries for information on both parameters and off-diagonal entries to account for the common part in their information. Generally, the results displayed in Table 3 are in accordance with the results for the 1-PL displayed in Table 2: Equations 5 and 6 produce comparable results and the precision of the diagonal approximations increases with sample size and decreases with test length. A difference is that the procedure using the inverse of the second order derivatives of the M-step does a very poor job for a sample size of 250.

TABLE 3

Approximation of standard errors for the 2-PL model magnitude relative to standard error, computed using complete observed information matrix

			Diagonal Observed Info Matrix		Complete Approximate Matrix		Diagonal Approximate Matrix		M-Step Approximate Matrix	
N	K	PL	α	β	α	β	α	β	α	β
250	5	1	0.921	0.905	0.936	0.955	0.903	0.902	0.553	0.915
		2	0.908	0.912	0.929	0.944	0.891	0.908	0.528	0.921
		3	0.897	0.911	0.934	0.937	0.886	0.909	0.539	0.907
	10	1	0.929	0.881	0.944	0.961	0.922	0.878	0.632	0.893
		2	0.941	0.901	0.966	0.956	0.948	0.901	0.654	0.908
		3	0.919	0.896	0.969	0.946	0.933	0.898	0.658	0.897
	20	1	0.885	0.853	0.904	0.953	0.879	0.853	0.592	0.859
		2	0.887	0.850	0.910	0.943	0.885	0.850	0.653	0.854
		3	0.867	0.844	0.934	0.963	0.885	0.848	0.615	0.842
	40	1	0.769	0.794	0.809	1.031	0.767	0.794	0.534	0.800
		2	0.779	0.793	0.817	0.975	0.776	0.794	0.605	0.799
		3	0.748	0.780	0.773	0.819	0.759	0.784	0.549	0.780
1,000	5	1	0.938	0.912	0.961	0.961	0.940	0.912	0.987	0.959
		2	0.932	0.923	0.957	0.960	0.930	0.923	0.976	0.967
		3	0.914	0.902	0.994	0.943	0.956	0.907	0.975	0.962
	10	1	0.960	0.912	0.970	0.977	0.961	0.912	0.759	0.928
		2	0.967	0.928	0.976	0.975	0.969	0.929	0.850	0.944
		3	0.952	0.924	0.968	0.970	0.956	0.924	0.853	0.948
	20	1	0.946	0.899	0.968	0.984	0.950	0.899	0.762	0.913
		2	0.948	0.898	0.967	0.987	0.950	0.900	0.856	0.914
		3	0.936	0.901	0.970	0.984	0.945	0.904	0.761	0.912
	40	1	0.921	0.910	0.945	1.029	0.921	0.911	0.719	0.922
		2	0.923	0.913	0.945	1.015	0.926	0.915	0.842	0.928
		3	0.907	0.893	1.143	1.283	0.912	0.894	0.693	0.899
4,000	5	1	0.942	0.922	0.957	0.964	0.940	0.922	0.985	0.964
		2	0.941	0.932	0.962	0.963	0.942	0.932	0.982	0.972
		3	0.939	0.897	1.090	0.954	1.049	0.907	0.991	0.947
	10	1	0.965	0.917	0.973	0.982	0.966	0.917	0.997	0.948
		2	0.971	0.931	0.975	0.979	0.971	0.931	0.998	0.956
		3	0.958	0.933	0.966	0.971	0.960	0.933	0.997	0.970
	20	1	0.961	0.911	0.976	0.998	0.960	0.911	0.986	0.935
		2	0.965	0.914	0.977	0.994	0.963	0.914	0.987	0.935
		3	0.951	0.906	1.000	1.001	0.969	0.910	0.983	0.937
	40	1	0.958	0.936	0.972	1.039	0.955	0.936	0.980	0.958
		2	0.961	0.939	0.975	1.031	0.958	0.939	0.980	0.958
		3	0.947	0.923	1.190	1.216	0.964	0.928	0.976	0.951
250	*	*	0.871	0.860	0.902	0.948	0.870	0.860	0.593	0.865
1,000	*	*	0.937	0.910	0.980	1.006	0.943	0.911	0.836	0.933
4,000	*	*	0.955	0.922	1.001	1.008	0.966	0.923	0.987	0.953
*	5	*	0.926	0.913	0.969	0.953	0.938	0.913	0.835	0.946
*	10	*	0.951	0.914	0.967	0.968	0.954	0.914	0.822	0.932
*	20	*	0.928	0.886	0.956	0.978	0.932	0.888	0.799	0.900
*	40	*	0.879	0.876	0.952	1.049	0.882	0.877	0.764	0.888

In Table 4, the results for the 3-PL model are given. In this case, item discrimination and difficulty parameters were estimated keeping the guessing parameter constant at its true value of 0.20. Further, the observed information matrix (Equation 5) was not computed because it was considered too complex. Therefore, Table 4 contains results for the block-diagonal approximate matrix (Equation 6) and the block-diagonal M-step matrix of second-order derivatives, both relative to the complete matrix defined by Equation 6. Results are not much different from those of the two studies reported above.

BEST COPY AVAILABLE

TABLE 4

Approximation of standard errors for the 3-PL model magnitude relative to standard error, computed using complete observed information matrix

			Diagonal Approximate Matrix		M-Step Approximate Matrix	
N	K	PL	α	β	α	β
250	5	1	0.911	0.894	0.968	0.950
		2	0.903	0.908	0.954	0.959
		3	0.864	0.873	1.129	1.075
	10	1	0.927	0.874	0.969	0.913
		2	0.942	0.905	0.969	0.932
		3	0.907	0.888	0.964	0.945
	20	1	0.886	0.856	0.912	0.882
		2	0.889	0.857	0.912	0.878
		3	0.851	0.835	0.914	0.897
	40	1	0.771	0.795	0.794	0.819
		2	0.780	0.793	0.801	0.815
		3	0.760	0.771	0.804	0.816
1,000	5	1	0.938	0.912	0.987	0.959
		2	0.932	0.923	0.976	0.967
		3	0.882	0.893	0.963	0.974
	10	1	0.960	0.912	0.991	0.942
		2	0.967	0.931	0.992	0.955
		3	0.930	0.905	0.991	0.964
	20	1	0.947	0.901	0.972	0.925
		2	0.949	0.903	0.973	0.925
		3	0.926	0.896	0.978	0.947
	40	1	0.922	0.910	0.944	0.932
		2	0.923	0.911	0.944	0.932
		3	0.905	0.888	0.949	0.932
4,000	5	1	0.942	0.922	0.985	0.964
		2	0.941	0.932	0.982	0.972
		3	0.898	0.907	0.970	0.979
	10	1	0.965	0.917	0.997	0.948
		2	0.971	0.931	0.998	0.956
		3	0.935	0.910	0.997	0.970
	20	1	0.961	0.911	0.986	0.935
		2	0.965	0.914	0.987	0.935
		3	0.943	0.912	0.993	0.960
	40	1	0.958	0.936	0.980	0.958
		2	0.961	0.939	0.980	0.958
		3	0.937	0.912	0.982	0.957
250	*	*	0.866	0.854	0.924	0.907
1,000	*	*	0.932	0.907	0.972	0.946
4,000	*	*	0.948	0.920	0.986	0.958
*	5	*	0.912	0.907	0.990	0.978
*	10	*	0.945	0.908	0.985	0.947
*	20	*	0.924	0.887	0.958	0.920
*	40	*	0.880	0.873	0.909	0.902

In general, all results give the impression that it must be possible to tune the CUSUM statistic using the block-diagonal approximation in such a way that a test with reasonable power characteristics can be obtained. The approach using the second order derivatives of the M-step is inferior to the other approximations and will not be further pursued. The block-diagonal approaches based on Equations 5 and 6 produced nearly identical results. However, the latter is easiest to compute. Therefore, the power of a CUSUM test based on this latter approach will be the topic of the next section.

The Power of the CUSUM Statistic

In this section, a simulation study of the power of a test based on a CUSUM statistic where the standard error is computed using the inverse of the block-diagonal of Equation 6 is presented. The model used was the 3-PL model with a fixed guessing parameter of 0.20. The item bank consisted of 100 items, with difficulties β equally spaced on the interval -1.00 to 1.00, and discrimination parameters α drawn from a log-normal distribution with mean zero and standard deviation equal to 0.10. Ability parameters were standard, normally distributed throughout the study. For the pretesting stage, four groups of 250 simulated test takers each were generated. Each group responded to 50 items: the first group responded to the items 1 to 50, the second group to 26 to 75, the third group to 51 to 100 and the fourth group to the items 1 to 25 and 76 to 100. So in the pretest design every item was presented to 500 respondents. Next, four batches, $g = 1, ..., 4$, of 1,000 adaptive-testing simulated test takers were generated. Every simulated test taker responded to 20 items. Person parameters were estimated by their posterior expectation using a standard normal prior, and item selection was by the maximum information criterium. Finally, CUSUM statistics $S_i(g)$, $g = 1, ..., 4$, were computed where the standard errors were inflated by 1.1 to correct for the bias identified in the previous section.

Consider the results displayed in Table 5. The first lines of the table relate to 10 replications without any misfitting items. The entries in the row labeled "fit" contain the proportions of item statistics $S_i(g)$, $g = 1, ..., 4$, with values lager than 2.5, 5.0, 7.5 and 10.0, respectively. So, for instance, the proportion of items with a value $S_i(4) > 5.0$ is equal to 0.04. The following rows contain the results of simulations where, in the online stage model, violations were imposed on every 5th item, that is, on item 5, 10, 15, 20, 25, ..., and so forth. So for every replication, 20 items were misfits while the remaining 80 items fitted the model. Again, 10 replications were made for every study. The model violations are listed in the rows labeled Model Violation. The entries $\gamma_i + 0.05$, $\gamma_i + 0.10$, and $\gamma_i + 0.20$ refer to studies where, for the misfitting items, the guessing parameter was augmented with 0.05, 0.10 and 0.20, respectively. The entries $\beta_i - 0.20$, $\beta_i - 0.40$, and $\beta_i - 0.60$ refer to studies where, for the misfitting items, the difficulty parameter was lowered with 0.20, 0.40 and 0.60, respectively. The rows labeled Misfit give the proportion of CUSUM statistics $S_i(g)$, $g = 1, ..., 4$, with values lager than 2.5, 5.0, 7.5, and 10.0 for the misfitting items. The rows labeled Fit give similar proportions for the fitting items. It can be seen that the proportions for the misfitting items are substantially larger than the proportions for the fitting items, the latter proportions are generally comparable to the proportions in the study without model violations. The study where misfitting items had their guessing parameter augmented from 0.20 to 0.25 seems to be an exception: here the proportion of fitting items with $S_i(g) > 2.5$ is equal to 0.21, so in this case the probability of a Type I error seems to be quite high. But generally, the proportion of Type I errors seems to be under control, while the proportion of misfitting items spotted is reasonable.

TABLE 5

Performance of CUSUM under various model violations

Timepoint	$g = 1$				$g = 2$				$g = 3$				$g = 4$			
Baseline without model violation																
$S_i(g) >$	2.500	5.000	7.500	10.000	2.500	5.000	7.500	10.000	2.500	5.000	7.500	10.000	2.500	5.000	7.500	10.000
Fit	0.030	0.000	0.000	0.000	0.040	0.010	0.000	0.000	0.070	0.020	0.000	0.000	0.080	0.040	0.010	0.000
Model violation $\gamma_i + 0.05$																
$S_i(g) >$	2.500	5.000	7.500	10.000	2.500	5.000	7.500	10.000	2.500	5.000	7.500	10.000	2.500	5.000	7.500	10.000
Misfit	0.000	0.000	0.000	0.000	0.200	0.100	0.000	0.000	0.400	0.100	0.000	0.000	0.350	0.100	0.050	0.000
Fit	0.000	0.000	0.000	0.000	0.010	0.000	0.000	0.000	0.030	0.000	0.000	0.000	0.050	0.010	0.000	0.000
Model violation $\gamma_i + 0.10$																
$S_i(g) >$	2.500	5.000	7.500	10.000	2.500	5.000	7.500	10.000	2.500	5.000	7.500	10.000	2.500	5.000	7.500	10.000
Misfit	0.250	0.000	0.000	0.000	0.500	0.100	0.000	0.000	0.550	0.400	0.100	0.000	0.650	0.450	0.300	0.050
Fit	0.010	0.000	0.000	0.000	0.040	0.010	0.000	0.000	0.060	0.010	0.000	0.000	0.090	0.030	0.010	0.010
Model violation $\gamma_i + 0.20$																
$S_i(g) >$	2.500	5.000	7.500	10.000	2.500	5.000	7.500	10.000	2.500	5.000	7.500	10.000	2.500	5.000	7.500	10.000
Misfit	0.150	0.050	0.000	0.000	0.250	0.100	0.050	0.000	0.300	0.100	0.050	0.050	0.450	0.150	0.050	0.050
Fit	0.050	0.000	0.000	0.000	0.130	0.030	0.000	0.000	0.170	0.040	0.010	0.000	0.210	0.090	0.030	0.000
Model violation $\beta_i - 0.20$																
$S_i(g) >$	2.500	5.000	7.500	10.000	2.500	5.000	7.500	10.000	2.500	5.000	7.500	10.000	2.500	5.000	7.500	10.000
Misfit	0.150	0.000	0.000	0.000	0.350	0.000	0.000	0.000	0.550	0.200	0.000	0.000	0.500	0.350	0.150	0.000
Fit	0.030	0.000	0.000	0.000	0.030	0.000	0.000	0.000	0.030	0.010	0.000	0.000	0.060	0.010	0.000	0.000
Model violation $\beta_i - 0.40$																
$S_i(g) >$	2.500	5.000	7.500	10.000	2.500	5.000	7.500	10.000	2.500	5.000	7.500	10.000	2.500	5.000	7.500	10.000
Misfit	0.300	0.100	0.000	0.000	0.750	0.400	0.150	0.000	0.900	0.600	0.400	0.150	0.850	0.800	0.600	0.400
Fit	0.030	0.000	0.000	0.000	0.040	0.000	0.000	0.000	0.060	0.000	0.000	0.000	0.090	0.040	0.000	0.000
Model violation $\beta_i - 0.60$	2.500	5.000	7.500	10.000	2.500	5.000	7.500	10.000	2.500	5.000	7.500	10.000	2.500	5.000	7.500	10.000
$S_i(g) >$	2.500	5.000	7.500	10.000	2.500	5.000	7.500	10.000	2.500	5.000	7.500	10.000	2.500	5.000	7.500	10.000
Misfit	0.200	0.050	0.000	0.000	0.750	0.250	0.050	0.050	0.950	0.600	0.200	0.050	1.000	0.750	0.450	0.150
Fit	0.030	0.000	0.000	0.000	0.060	0.000	0.000	0.000	0.100	0.000	0.000	0.000	0.100	0.040	0.000	0.000

Conclusion

In this study it was shown that standard errors can be reasonably approximated using a block-diagonal approximation to the observed Fisher information matrix. Further, it was shown that the power of a test based on a CUSUM statistic using this approximated standard error is well under control. Tuning the CUSUM statistic to a specific application can be done by running a simulation study as in the previous section, with test administration design features as they apply in this specific application, and evaluate how the underestimated standard errors must be translated into values of k and h to obtain acceptable power characteristics.

References

Efron B. and Hinkley, G. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika, 65,* 457–482.

Glas, C. A. W. (1992). A Rasch model with a multivariate distribution of ability. In M. Wilson, (Ed.), *Objective measurement: Theory into practice, Vol. 1* (pp.236–25). Westport, CT: Ablex Publishing Corporation.

Glas, C. A. W. (1998a). Detection of differential item functioning using Lagrange multiplier tests. *Statistica Sinica, 8,* (3), 647–667.

Glas, C. A. W. (1998b). *Modification indices for the 2-PL and nominal response model* (Research Report 98-04). Enschede, The Netherlands: University of Twente, Department of Educational Measurement and Data Analysis.

Glas, C. A. W. (1998c). *Quality control of online calibration in computerized adaptive testing* (Research Report 98-03). Enschede, The Netherlands: University of Twente, Department of Educational Measurement and Data Analysis.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B, 44,* 226–233.

Mislevy, R. J. (1986). Bayes model estimation in item response models. *Psychometrika, 51,* 177–195.

Veerkamp, W. J. J. (1996). *Statistical methods for computerized adaptive testing.* Unpublished doctoral thesis, Twente University, Enschede, The Netherlands.

BEST COPY AVAILABLE

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").